edWisor

Predictive Analytics Using R and Python

- MUQUAYYAR AHMED
DATA SCIENTIST

We learnt!!

edWisor

1. Introduction to Predictive Analytics

- Skills Needed
- What is Data Analytics
- Applications
- Type of Problems
- CRISP-DM Process

2. Introduction to R

3. Introduction to Python

4. Missing Value Analysis

5. Outlier Analysis

Today's Agenda

- Feature Selection

Variable Importance/Feature Selection

- What is Feature Selection??
  - Selecting a subset of relevant features (variables, predictors) for use in model construction
  - subset of a learning algorithm's input variables upon which it should focus attention, while ignoring the rest

- Dimensionality reduction

- Curse of dimensionality!

Correlation Analysis

- Correlation tells you the association between two continuous variables
- Ranges from -1 to 1
- Measures the direction and strength of the linear relationship between two quantitative variables
- Represented by “r”
- Correlation can be calculated as

$$r = \frac{Covariance\ (x,y)}{S.D.(x)S.D.(y)}$$

- Covariance

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$

Chi-Square test of Independence

- Compares two variables in a contingency table to see if they are related
- Hypothesis Testing
  - Null Hypo: Two variables are independent
  - Alternate Hypo: Two variables are not independent
- Uses contingency table for better representation

- Chi-square test can be calculated as

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Contd..

- Degrees of Freedom

$$(\text{number of rows} - 1)(\text{number of columns} - 1)$$

- Calculate critical value using table

- If Chi-square statistic is greater than Critical value then reject null hypothesis

edWisor

Statistical Techniques & Algorithm!!!

- Principal Component Analysis (PCA)
- Singular Value Decomposition (SVD)
- Random Forest
- Correlation Analysis
- Chi-square Test
- ANOVA
- “Fselector” package in R